The $100 Genome: Implications for the DoD

MITRE

The $100 Genome: Implications for the DoD

Contact: D, McMorrow - dmcmorrow@mitre.org

December 2010

JSR-10-100

JASON
The MITRE Corporation
7515 Colshire Drive
McLean, Virginia 22102-7508
(703) 983-6997

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
December 15, 2010	Technical	

4. TITLE AND SUBTITLE

The $100 Genome: Implications for the DoD

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER
13109022

5e. TASK NUMBER
PS

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The MITRE Corporation
JASON Program Office
7515 Colshire Drive
McLean, Virginia 22102

8. PERFORMING ORGANIZATION REPORT NUMBER

JSR-10-100

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

OSD/AT&L/DDR&E/PD
Pentagon, Rm 38854
Washington, DC 20520

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Rapid advances in DNA sequencing and other technologies are ushering in an era of personal genomics. Soon it will be possible for every individual to have access to the complete DNA sequence of his or her genome for a modest cost. This development, coupled with the improving ability to predict how genetic variation affects susceptibility to disease, response to medical treatment, and other important phenotypes, will have a transformative effect on health care. This will be far reaching in civilian medical practice, and it could be used in the assessment of personnel at all stages of their military service.

JASON was asked to consider the impact of anticipated advances in genome sequencing technology over the next decade, and to assess the relevant operational opportunities and challenges that will be presented by these technologies.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Melissa Flagg
					19b. TELEPHONE NUMBER *(include area code)* 703-614-9443

Contents

1 EXECUTIVE SUMMARY

Introduction

Rapid advances in DNA sequencing and other technologies are ushering in an era of personal genomics. Soon it will be possible for every individual to have access to the complete DNA sequence of his or her genome for a modest cost. This development, coupled with the improving ability to predict how genetic variation affects susceptibility to disease, response to medical treatment, and other important phenotypes, will have a transformative effect on health care. This will be far reaching in civilian medical practice, and it could be used in the assessment of personnel at all stages of their military service.

The U.S. military is a major consumer of medical services and has special medical needs compared to the general population. As the revolution in personal genomics proceeds, the military stands to benefit by implementing genomic technologies that enhance medical status and improve treatment outcomes. Furthermore, both offensive and defensive military operations may be impacted by the applications of personal genomics technologies through enhancement of the health, readiness, and performance of military personnel. It may be beneficial to know the genetic identities of an adversary and, conversely, to prevent an adversary from accessing the genetic identities of U.S. military personnel.

Study Charge

JASON was asked to consider the impact of anticipated advances in genome sequencing technology over the next decade, and to assess the relevant operational opportunities and challenges that will be presented by these technologies. The specific questions we sought to address in the study were:

1. What types of genetic information are likely to be most informative in personalized medicine relevant to military personnel?
2. What types of genetic information are likely to have tactical benefit in either offensive or defensive military operations?
3. What are the capabilities and costs of current technologies for assaying genetic variation, and are these capabilities sufficient to acquire the desired genetic information at reasonable cost?

4. How do recent research findings on the relative importance of different types of variation in human disease (e.g. common variants vs. rare variants; single-nucleotide polymorphisms vs. copy-number changes) affect the assessment of desired information and the technology needed to acquire it?
5. What new genomic technologies are emerging that are likely to benefit the military in the near term, and what is the likely roadmap for advances in genome sequencing technology over the next decade or so?
6. What capabilities will be required to analyze large sets of genomic data and accompanying phenotype data to derive information regarding the genetic basis for phenotypes of unique concern for force protection?
7. How should personal genomic information be handled to maintain the security of that information?

Summary

The first draft sequences of the human genome were published a decade ago at a cost of ~$300M. Although these data provided an unprecedented view of the genetic blueprint of humans, the prohibitive cost of DNA sequencing made difficult the correlation of genetic variations with specific traits. Successive improvements in "second-generation" DNA sequencing platforms over the last five years reduced the cost of sequencing by approximately an order of magnitude each year. An entire human genome can now be sequenced in a matter of days for a retail cost of $20,000, and "third-generation" DNA sequencing systems soon to be released will drive costs of reagents to below $100, although machines, labor and data processing expenses will add to the cost of each genome.

Given technological developments, we believe that DNA sequencing costs will no longer be a factor limiting personal human genomics technologies. Therefore DNA sequencing information is likely to be broadly applied to health assessment, therapeutic decisions, and predicting phenotypes of interest. Although current understanding of the linkages between the genotypes of individuals and their phenotypes is limited, researchers are pursuing the challenging topic of linking genetic variations to specific diseases or other traits and soon will have an enormous amount of DNA sequence data available to drive these linkage studies. Furthermore, individuals will have ready access to their own genome sequences at modest cost,

and these data can be searched for genetic indicators of the propensity for disease traits, to the extent such linkages can be established.

The DoD is well positioned to capitalize on personal genomics technologies and could choose to be full partners with industry and academic leaders in this field. The DoD has a large, well-defined population in generally good health, together with their medical health records, which could facilitate valuable longitudinal studies correlating genotype and phenotype. The existing military health care delivery system is in a position to be adapted to accommodate personalized genomics information. The DoD can leverage many of the advances within the civilian healthcare system, but there are also particular phenotypes that are of interest to the military that are unlikely to be high-priority goals for research and development in the civilian sector, and which therefore merit special DoD attention.

Findings

The technologies for DNA sequencing and personal genomics are advancing at a rapid pace. The cost of obtaining personal genomics data will continue to fall and this information will be used to make predictions regarding health and performance. Soon, researchers will have an enormous amount of data to use in the challenging effort to link genotypes with phenotypes of interest. Correlating genotypic markers with phenotypic traits will be challenging, particularly for traits controlled by multiple genes with low penetrance, and will require the development and application of powerful bioinformatics tools: this represents the subject of an ongoing research effort. Furthermore, there are genetic features beyond the genome sequence that stand to complicate efforts to link genotypes with phenotypes, and yet promise opportunities for understanding physiology and disease. Both epigenetics and the human microbiome, for example, are known to exert major effects on human phenotypes.

The specific findings of the study are the following:

1. The $100 genome is nearly upon us, and soon the cost of DNA sequencing will no longer be a limiting factor in genomic analysis.
2. The era of personal genomics has already begun, but the practical application of genomic information has thus far been limited.
3. Broader application of genetic information will require deeper knowledge of genotype-phenotype correlations, a subject of substantial, ongoing research.

4. Many phenotypes of relevance to the DoD are likely to have a strong genetic component, for which better understanding may lead to improved military capabilities.
5. Certain phenotypes will also depend upon epigenomic and microbiomic contributions. However, human epigenomes and microbiomes are diverse and will change with time, and therefore complete datasets for these genetic signatures cannot be collected.
6. The DoD already maintains a comprehensive medical database for its personnel that eventually will also include their complete genome sequences.
7. The DoD will benefit by organizing personnel data into phenotypes of relevance to the military, then correlating those phenotypes with genetic information.

Recommendations

The DoD can benefit significantly by employing personal genomics technologies when evaluating the health and performance characteristics of their personnel. The DoD could take a leading role in the personal genomics era, and become full partners with industry and academia in creating useful information from genotype and phenotype data. Alternatively, the DoD could choose to play a more limited role in the research necessary to link genotypes with phenotypes, and pursue only those aspects that are of special interest to the military and that would otherwise not be pursued by the civilian sector.

The DoD can harness the advances in personal genomics technology by taking the actions described below.

Major Recommendation

The DoD should establish policies that result in the collection of genotype and phenotype data, the application of bioinformatics tools to support the health and effectiveness of military personnel, and the resolution of ethical and social issues that arise from these activities.

Specific Recommendations

DoD Military Health System

1. Establish procedures for the collection and archiving from all military personnel DNA samples that are compatible with subsequent genotype determination.
2. Plan for the eventual collection of complete human genome sequence data from all military personnel.

3. Arrange for the secure, long-term storage of DNA sequence data.
4. Prepare for the collection of epigenome and microbiome data when appropriate.

DoD Office of Health Affairs

1. Determine which phenotypes are of greatest relevance to the DoD.
2. Cooperate with health care professionals to collect and store these data.
3. Use bioinformatics tools to correlate genetic information with phenotypes to discover linkages between the two datasets that will ultimately allow genotype information to be used productively.

JASON was pleased to conduct this study, which allowed us to assess the impact that rapidly-improving DNA sequencing technologies will have on genomics and the implications of these advances for the DoD. DNA sequencing costs, which previously had been one of the limiting factors on the broad use of personal genomics, will soon become comparable to most routine medical testing. The changing economics will enable advances in correlating genotype and phenotype that stand to benefit the DoD. We also anticipate that the added layers of genomic complexity derived from epigenetics and human microbiomics will offer additional opportunities for improving the health and enhancing the performance of military personnel.

2 INTRODUCTION

In this report we summarize our considerations and findings of the 2010 JASON Summer Study entitled "The $100 Genome: Implications for the DOD". The study charge from DoD DDR&E was to address the following questions:

1. What types of genetic information are likely to be most informative in personalized medicine relevant to military personnel?
2. What types of genetic information are likely to have tactical benefit in either offensive or defensive military operations?
3. What are the capabilities and costs of current technologies for assaying genetic variation, and are these capabilities sufficient to acquire the desired genetic information at reasonable cost?
4. How do recent research findings on the relative importance of different types of variation in human disease (e.g. common variants vs. rare variants; single-nucleotide polymorphisms vs. copy-number changes) affect the assessment of desired information and the technology needed to acquire it?
5. What new genomic technologies are emerging that are likely to benefit the military in the near term, and what is the likely roadmap for advances in genome sequencing technology over the next decade or so?
6. What capabilities will be required to analyze large sets of genomic data and accompanying phenotype data to derive information regarding the genetic basis for phenotypes of unique concern for force protection?
7. How should personal genomic information be handled to maintain the security of that information?

These questions become particularly important given the striking advances in DNA sequencing technologies that have reduced both the expense and the time needed to collect the near-complete DNA sequence of an entire human genome. This capability will provide researchers with a considerable increase in data important for correlating genetic differences with disease and other phenotypic traits. Furthermore, this capability will provide medical personnel with a detailed map of the genetic distinctions of individual patients, which may be used to make diagnostic and therapeutic decisions.

2.1 Scope of the Study

To address the questions from our study charge, we first considered the technical advances in DNA sequencing technology. We were particularly interested in assessing whether future DNA sequencing platforms could deliver DNA sequence data at a cost and speed that will enable personal genome sequences to be collected for all military personnel, and with the accuracy necessary for medical applications. We examined the current utility of this data for making phenotypic predictions, evaluated research models for correlating genetic markers with phenotypes, and considered biological mechanisms other than DNA sequence that may confound attempts to predict phenotypes based solely on DNA sequence data. Particular attention was given to the human microbiome, which is large, complex, and can have important influences on normal human function and disease.

Our efforts were aided by the following briefers who we thank for their helpful insights and discussion:

DNA Sequencing Technologies

Eric Schadt – Chief Scientific Officer, Pacific Biosciences
Jonathan Rothberg – Founder and Chief Executive Officer, Ion Torrent
Omead Ostadan – Vice President of Marketing, Illumina
Jeff Schloss – National Institutes of Health

Genotype/Phenotype Correlations

David Galas – Senior Vice President for Strategic Partnerships, Institute for Systems Biology
David Haussler – Professor, University of California, Santa Cruz
David Altshuler – Professor, Harvard Medical School and Broad Institute

Epigenetics

Barbara Wold – Professor, CalTech

Human Microbiome

Rob Knight – Assistant Professor, University of Colorado

Personal Genomics

Anne Wojcicki – 23andMe

DoD Applications

Randall Kincaid – Scientific Director, TMT, Defense Threat Reduction Agency (DTRA)

2.2 The Challenge of Sequencing and Assembling Human Genomes

The utility of personal genomics will rely in part on the collection of genetic data that can be correlated with human traits. There will be many genetic markers such as common single-nucleotide polymorphisms (SNPs), gene deletions and gene duplications that will be indicative of phenotypic outcomes. However, many rare SNPs and other types of genetic variations will also be informative. Therefore, recording an individual's entire genomic DNA sequence provides the maximal amount of genetic information compared to less-comprehensive genetic tests such as DNA arrays.

The past and current state of DNA sequencing technologies were assessed to determine whether the efficiency of whole-genome DNA sequencing will progress to a point where it is cost-effective for all individuals to have their genome sequence data collected. Various aspects of human genome sequencing efforts are discussed below.

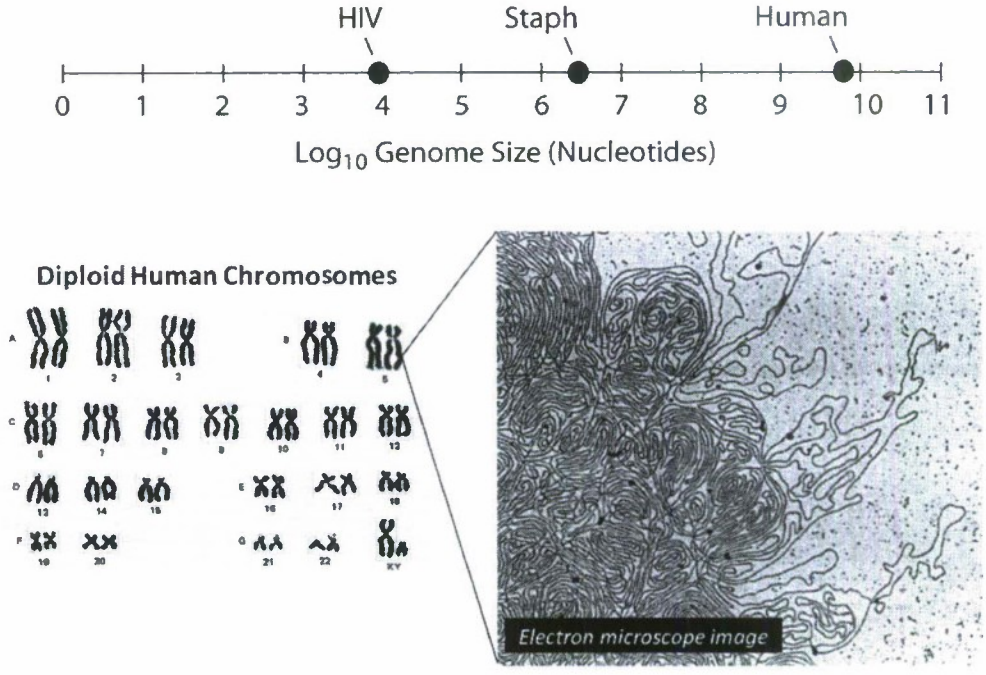


Figure 1. Size and complexity of human genomic DNA. Top: Number of DNA base pairs in the genomes of human immunodeficiency virus (HIV), *Staphylococcus aureus* (Staph), and the nuclear chromosomes of human cells. Bottom: The 23 pairs of human chromosomes carry the approximately 3 billion base pairs of nuclear DNA in each cell. The electron microscope image includes about 1 million DNA base pairs that have become unspooled from a chromosome.

Human Genomes and the DNA Sequencing Challenge

The challenge faced by those who wish to generate a complete DNA sequence of an individual's genome is great. The complete set of unique DNA sequences in each diploid cell is distributed between mitochondrial genomes (just over 16,500 base pairs in each copy) and 23 pairs of chromosomes (approximately 6 billion base pairs) (**Fig. 1**). Each chromosome is not sequenced as an individual unit, but rather all chromosomes from many cells are randomly fragmented into pieces ranging from ~100 to ~1000 nucleotides in length and these fragments are then sequenced. These DNA sequence "reads" are then reassembled into their original genetic context using computer algorithms that identify overlapping identity between millions of sequenced fragments. As noted later, this sequencing and reassembly strategy creates challenges when working to obtain a complete and error-free genome sequence.

	Sanger Sequencing	454 Life Sciences	Illumina	Applied Biosystems
DNA isolation method	Bacterial clone	PCR on beads	PCR colonies	PCR on beads
Paralellize method	Capillary array	Microfab plate	Surface clusters	Magnetic surface
Sequencing method	Dideoxy chain termination	PPi-mediated fluorescence	Fluorescent dNTP incorporation	Fluorescent oligo incorporation
Read length (bp)	800	400	100	50
Reads per run	384	1,000,000	200,000,000	1,000,000,000
Throughput	300 kb / hr	0.5 Gb / 10 hr	30 Gb / 10 days	50 Gb / 14 days

Figure 2. Comparison of the major forms of DNA sequencing platforms used over the last two decades.

DNA Sequencing History

For two decades following the mid 1970s, nearly all DNA sequence data was collected by using either Maxam-Gilbert DNA sequencing (selective chemical modification) (Maxam and Gilbert, 1977) or Sanger sequencing (Sequence by Synthesis [SBS] chain termination) (Sanger and Coulson, 1975) methods (**Fig. 2**). Both methods, sometimes called first-generation sequencing systems, involved extensive preparative work and allowed an individual research to sequence ~100 nucleotides per sample per day. Due to technical constraints of the Maxam-Gilbert sequencing and due to the accuracy of DNA sequence reads from the Sanger method, the

latter method became more widely used by individual research laboratories. However, the manual nature of the protocol and the lower-resolution slab gel separation limited Sanger sequencing typically only to the determination of gene-sized short stretches of genome sequence.

By the 1990s, portions of the Sanger protocol were being automated and capillary electrophoresis with fluorescently-labeled chain terminators allowed longer read lengths and faster analyses to be achieved. Dozens of samples could be run in parallel, and dozens of machines running constantly were used to sequence large portions of genomic DNA from many species. Partial automation of the Sanger protocol enabled a single instrument to sequence hundreds of thousands of DNA nucleotides per day. This capability was an important component of the molecular biology revolution of the late 20th century. Furthermore, instead of sequencing specific purified clones of genome fragments, researchers expedited the process by sequencing random fragments of the target genome, and used computer algorithms to reassemble the short reads into contiguous genome sequences. These technical advances facilitated the sequencing of entire bacterial genomes, and eventually were used to produce the first near-complete human genome sequences. These same technical aspects are still of great interest to those seeking to optimize DNA sequencing methods to create faster, cheaper, more compact, and more accurate DNA sequence data collection systems.

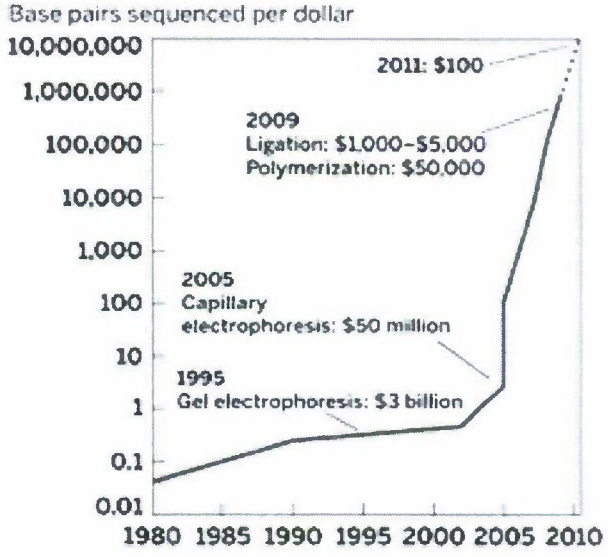


Figure 3. The improvements in DNA sequencing efficiency over time. Costs excludes equipment and personnel. Figure adapted from that published previously (Arnaud 2009).

Perhaps most importantly, each technical advance has led to exponential reductions in the per-base-pair cost of sequencing DNA (**Fig. 3**). A comparison is commonly made between this trend and "Moore's Law" that predicted a doubling of transistor density every 2 years. This type of exponential improvement trend can be observed for many other technological areas spanning from modes of transportation to the density of energy storage in batteries. Indeed, DNA sequencing technologies have advanced at a pace far greater than Moore's Law for transistor density, so that it is now possible to order your personal genome sequenced today for a *retail cost* of under ~$20,000. This cost will likely fall to less than $1,000 by 2012, and to $100 by 2013.

At costs below $1,000 per genome, a number of intriguing applications of DNA sequencing become cost effective. For example, researchers will have access to thousands or even millions of human genomes to seek correlations between genotypes and phenotypes. Medical doctors will be able to order genome sequencing along with the standard laboratory tests, and will likely do so if they believe that knowledge of the DNA sequence will facilitate patient diagnosis and/or treatment. Even web-based genetic testing service companies will exploit full genome sequences to gather and dispense medical and ancestry information, and provide genetic counseling. We also see tremendous value for DoD and VA interests, which is the main focus of this study.

3 DNA SEQUENCING TECHNOLOGY ASSESSMENT AND IMPLEMENTATION

JASON critically assessed the existing and emerging DNA sequence technologies to determine whether routine personalized genome analyses can be realized. It is clear that rapid improvements have been made to give rise to existing DNA sequencing systems and that it is reasonable to assume that improvements will continue to be made at a similar pace. Below are described some of the key technological aspects of DNA sequencing systems that are bringing forth a revolution in personalized genomics.

3.1 Current DNA Sequencing Technologies

Personalized genomics can be greatly facilitated by the ability to rapidly sequence human genomes with high accuracy and with low cost. As described below, current DNA sequencing platforms (**Fig. 3**), sometimes described as second-generation DNA sequencing technologies, have strived to address key technological factors that limit accuracy and cost, and performance (Schadt et al., 2010). In a number of areas, technological advancements have entered a range wherein personalized genomics is viable. Commercial DNA sequencing systems currently used to sequence entire human genomes are listed below along with their key features.

Machine	**Technology/Features**
454/Roche	DNA colonies on beads sequenced by fluorescent monomer incorporation
Solexa/Illumina	DNA colonies on a surface sequenced by fluorescent monomer incorporation
Helicos	Single DNA molecules on a surface sequenced by fluorescent monomer incorporation
ABI SOLiD	DNA colonies on beads attached to a surface and sequenced by fluorescence oligomer incorporation
Dover Polonator	DNA colonies on beads attached to a surface and sequenced by fluorescence oligomer incorporation

Sequencing speed. Nearly all commercial DNA sequencers operate via sequence by synthesis (SBS) technologies, wherein the selective addition of a single nucleotide (or the selective ligation of a short oligonucleotide) is required to read out the identity of the nucleotide being evaluated. Although these reactions typically occur on a sub-second or a sub-millisecond timescale, the systems require stepwise reagent additions and washes that occur on a timescale of seconds due to fluidics limitations. However, these reduced timescales also allow data acquisition to occur with reduced probability of error.

These slow speeds mean that it is not possible to sequence long stretches of DNA with existing technologies. Therefore sequencing the millions of nucleotides of a typical bacterium, or the billions of a human in the context of intact chromosomes is not achievable. To overcome this problem, existing commercial DNA sequencing platforms do not sequence entire chromosomes intact, but rather sequence very short fragments (~30 to ~1000 nucleotides) in a massively parallel fashion, and then use computer algorithms to reassemble these fragments to yield to the genomic sequence DNA within its proper linkage context. The genomic DNA is fragmented, the short DNAs are uniquely distributed to different reaction wells or locals, and these are each sequenced in place.

Sequencing Accuracy. The optimal DNA sequence data set for any organism is a 100% accurate collection of nucleotide sequences that are assembled into their appropriate chromosomal units. However, practical considerations necessary to drive sequencing speed higher and costs lower can cause compromises in the completeness and accuracy of datasets. Errors usually enter this process due to misreading the output from the sequencing reactions or by errors in reassembling the sequence fragments into contiguous chromosomal units. A system that sequences single molecules may reduce reagent cost and increase sequencing speed, but also may be more prone to errors in the sequence reaction event (e.g. an error by DNA polymerase) compared to methods that sequence clusters of identical DNAs in a reaction well.

Errors caused by the data recording instrumentation or by the inability of the analysis software to make the correct nucleotide call will introduce errors regardless of the type of sequencing strategy used. These errors can be reduced dramatically by collecting DNA sequence data far in excess of the number of nucleotides in the organism's genome. For example, assembled genome sequences are typically done with ~30-fold sequencing coverage, which means that each nucleotide has on average been sequenced 30 times. Since the fragments are

randomly generated and randomly sequenced, there are some nucleotides that have greater sequencing coverage than others.

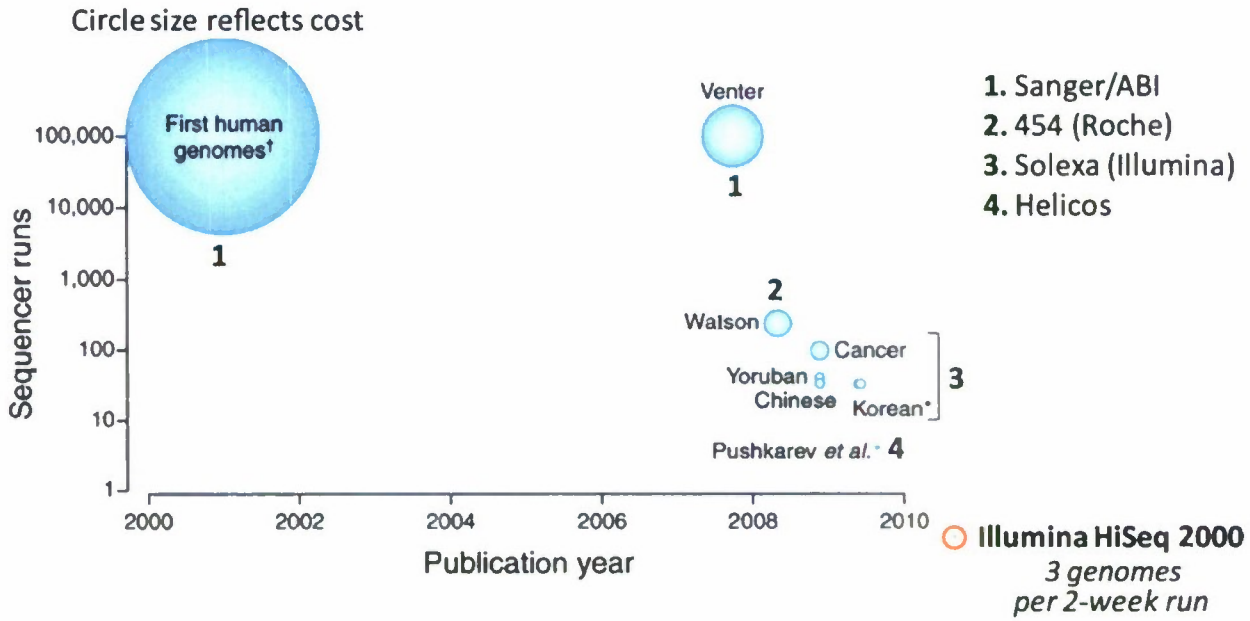


Figure 4. Reductions in the cost of sequencing complete human genomes over time. Examples given are for published human genomes with the exception of the Illumina HiSeq 2000 which was recently released. Size of circle represents the relative cost, where Venter equals $1M. Red circle for the Illumina HiSeq 2000 encompasses the actual data point, which reflects the characteristics of the system (2 to 3 human genomes per run with a wholesale reagent cost of ~$1,000 per genome). Graphic was adapted from that published by Li and Wang (2009).

The first bacterial genome sequence datasets included considerable inaccuracy, with some genome segments carrying approximately 4% erroneous nucleotide assignments. Even far lower error rates can cause considerable problems when conducting bioinformatics analyses. Although improving technologies and strategies cannot completely exclude error, the frequency of errors can be reduced to levels that are manageable with the current commercial systems.

Sequencing Cost. The number of nucleotides that can be sequenced per dollar is rising at an exponential rate (**Fig. 3**). Thus, the cost of sequencing complete human genomes is falling dramatically, with a rate of ~30 fold reduction in cost per year over the last six years (**Fig. 4**). These striking reductions in cost have been made on earlier SBS sequencing platforms largely by increasing the parallelism of the sequencing reactions (via reducing feature density or size) and by optimizing reactions for longer reads. Interestingly, the reductions in cost and read length realized by current sequencing systems allow the collection of redundant sequencing reads of greater length, which reduces the frequency of errors in the final datasets.

3.2 Future DNA Sequencing Technologies

Numerous companies are working on new technologies, called third-generation sequencing systems, that may contribute to further improvements in the parameters discussed above. The technologies of three companies were examined in some detail to assess the potential for enhanced DNA sequencing speed, accuracy and efficiency, although additional technologies are also being pursued (Schadt et al., 2010).

Company	**Technology/Features**
Pacific Biosciences	Zero mode waveguide sequencing of single DNA molecules using immobilized DNA polymerase and fluorescently-tagged nucleotides.
Ion Torrent	DNA colonies on beads sequenced using CMOS-based pH sensing of nucleotide additions.
Oxford Nanopore	Single DNA molecule sequencing using conductance changes across nanopore-studded membranes.

Pacific Biosciences: Single-molecule DNA sequencing by zero mode waveguide technology. Pacific Biosciences is developing sequencing technology that overcomes the need for synchronized reagent addition, which should reduce reagent costs, increase read lengths, and dramatically reduce the time needed to sequence each nucleotide. The key features of the sequencing platform is the use of zero mode waveguide technology to selectively image fluorescent mononucleotides as they are added to an elongating DNA chain by DNA polymerase. A laser illuminates only the lower third of each reaction well (~100 nm in diameter), which contains ~ 20 x 10^{-21} liters including a single immobilized DNA polymerase and associated DNA. This arrangement allows the determination of which distinctly-labeled nucleotide type has been bound by DNA polymerase just before its addition to the growing DNA chain, since unbound nucleotides spend very short times diffusing through this region of the reaction well (**Fig. 5**).

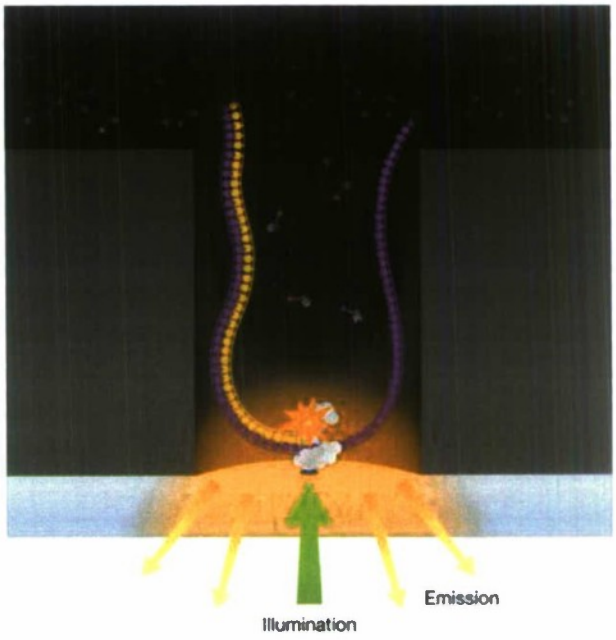


Figure 5. Reaction well of a Pacific Biosciences DNA sequencing system. Zero mode waveguide technology allows laser illumination of the lower portion of the reaction well, which allows imaging of a single fluorescent nucleotide that is retained in the active site of DNA polymerase. Unbound nucleotides quickly diffuse in and out of the illuminated region.

Theoretically, this system should allow very long DNA sequences to be assessed at the speed corresponding to the rate for nucleotide incorporation by DNA polymerase. However, camera technology is currently insufficient to keep pace with the burst speed of DNA polymerases (~50 to 1000 nucleotides per second). Therefore, the polymerization reaction is slowed to yield 1 to 5 nucleotides per second. Other problems include inactivation of DNA polymerase (e.g. via denaturation or laser-induced destruction) that limit the read length. However, improvements in these hardware and biological components are to be expected, and single-molecule sequencing using this method should yield considerable improvements over second-generation sequencing systems. Representatives at Pacific Biosciences project that by 2014, a system will be operational that collects a 300 gigabase dataset collection within 15 minute runs with each of 160,000 wells operating at 50 nucleotides per second read speed.

Ion Torrent: DNA sequencing by CMOS-based pH sensing. Ion Torrent has developed DNA sequencing chips based on CMOS (complementary metal-oxide-semiconductor) technology. This system uses multiplexed and miniaturized ion sensors to detect the release of a hydrogen ion that results from the addition a DNA nucleotide to the growing DNA chain during synthesis (**Fig. 6**). This detection technology avoids the use of chemically modified reagents (e.g.

fluorescent nucleotides), but retains the need for synchronized synthesis on clusters of DNAs to generate sufficient change in pH to reliably detect nucleotide additions.

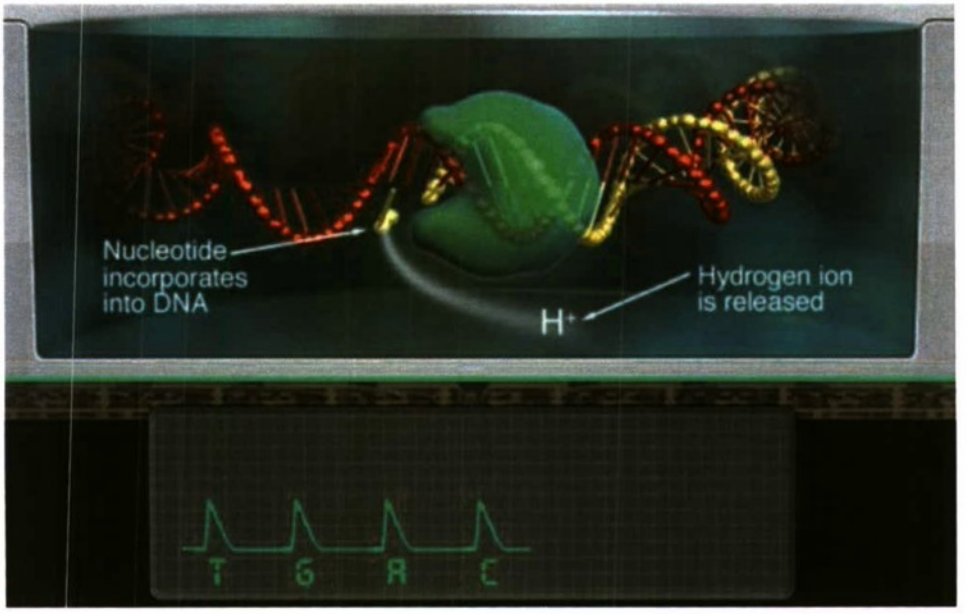


Figure 6. Schematic of the DNA polymerization event that releases a proton. Clusters of identical DNAs undergoing synchronized synthesis temporarily lower pH of the reaction well, which is monitored by individual ion sensors.

Reductions in reagent cost coupled with the relatively inexpensive cost of the machine mean the overall cost of sequencing is reduced substantially over existing second-generation systems. Initial Ion Torrent chips carry feature sizes of 350 nm, 23 million sensors per chip, and allow 30-fold sequencing coverage of a human genome for under $6,500. However, these chips are made using chip fabrication facilities constructed in 1995. Dramatic reductions in feature size and density can be achieved simply by using more recent chip fabrication facilities, which effectively leverages the investments made to improve computer chip feature density to create massive improvements in DNA sequencing capability. Therefore, DNA sequencing chips that permit complete collection of a human genome for less than $100 seems within easy reach.

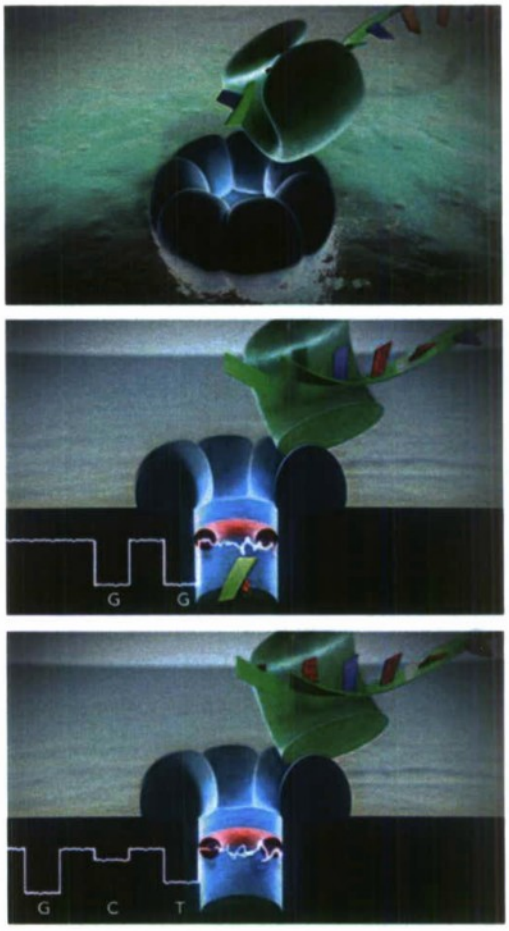


Figure 7. DNA sequencing using nanopores. In one possible architecture, a DNA to be sequenced is successively cleaved by a nuclease (top) and the released nucleotide enters a protein nanopore (middle). Temporary blockage of the pore changes current flow in a manner distinct to each nucleotide type, which is recorded and converted into nucleotide sequence data (bottom).

Oxford Nanopore: DNA sequencing by nanopore-based DNA threading. A less-mature method for DNA sequencing involves the use of a protein nanopore and associated protein and surface components to evaluate the sequence of a single DNA molecule (Clarke et al., 2009). In one system envisioned, a nuclease protein is fused to a hemolysin protein pore that penetrates a lipid or silicon barrier (**Fig. 7**). A single DNA strand is bound by the nuclease, which processively cleaves and releases single nucleotides. These nucleotides then travel through the pore and are detected by recording the electrical current that flows through the pore. The release of specific nucleotides are successively determined based on the distinct pore currents and the dwell times associated with the passage of each nucleotide type through the adjoining nanopore.

Multiplex arrangement of nuclease/pore complexes and the associated sensors could give rise to an efficient DNA sequencing system that uses no specialized reagents. However, the

biological-derived components of the system likely will require considerable molecular engineering to perform the tasks with the right characteristics necessary for the system to compete with other sequencing technologies. Also, creating large multiplexed arrays of pores and their associated detectors will require technological advances. Additional variations of nanopore DNA sequencing are being pursued, but likewise, numerous technical challenges must be overcome for this general approach to become practical.

3.3 Who is Buying DNA Sequencing Machines and What are Their Goals?

The market for DNA sequencing hardware was $480M in 2008, and sales are expected to increase approximately 20% annually through 2013. These machines can cost up to $750K per unit, which restricts their sales to well-funded academic laboratories (single units) or to university core facilities, research institutes, and pharmaceutical companies (multiple units). In addition to permitting full genome sequencing to be achieved, these machines are allowing researchers to conduct new types of biological studies that demand high-throughput DNA sequencing data. Such studies commonly involve the analysis of the collection of transcripts that are made by cells or the collection of DNAs that may be bound by specific transcription factors. An example of a use in drug discovery is the resequencing of the genomes of bacteria that have become resistant to antibiotics.

In addition to the typical users, there have been several high-profile large-scale purchases of machines by individual institutes. The Broad Institute in Cambridge MA has recently announced the purchase of 51 Illumina HiSeq 2000 machines (combined list price of $38 M). Each Illumina HiSeq 2000 machine can sequence the equivalent of 2 to 3 human genomes for each 10-day run, to give a minimum of 73 genomes per year at an approximate wholesale reagent cost approaching $1,000 per genome.

The Broad Institute services the DNA sequencing needs of their associated scientific community, and so some of this new sequencing capacity will be used for this purpose. However, the Institute and their collaborators also have access to ~200,000 human DNA samples that are coupled with well-established phenotypic data, and they plan to begin sequencing these samples to correlate genotypes with phenotypes. However, it is noteworthy that 51 HiSeq 2000 machines operating continuously can sequence almost 4000 genomes per year, but this still would require 50 years to sequence all of the samples currently held by the Broad Institute. Therefore it is certain that third-generation sequencing machines will be required to complete the sequencing of these samples in a reasonable timeframe.

A remarkable purchase of 128 Illumina HiSeq 2000 machines (~$95M) has recently been made by the former Beijing Genome Institute (now called BGI), located in Hong Kong. Reports suggest they have various plans for these machines, which have the capacity to sequence ~8,000 human genomes per year. This private institute used bank loans and grants to purchase these sequencers, and they are striving to turn a profit with their DNA sequencing operations. To this end, they are conducting a fee-for-service business with collaborators around the globe. For example, they previously collaborated with researchers at the University of Copenhagen to sequence the genome of a 4,000-year-old frozen man from Greenland, and split the ~$500,000 cost. In addition to this activity, they are conducting internal research programs to sequence and analyze 10,000 bacterial genomes, 1,000 plant genomes, and numerous human genomes (Cyranoski 2010). A particularly noteworthy project they have publically discussed involves the sequencing of 2,000 school children to look for markers that correlate with educational test scores.

Given that DNA sequencing costs are dropping by ~30 fold per year, the advantage gained by investing heavily in one existing DNA sequencing technology is unclear. Although the Broad Institute and BGI are currently among the world leaders in DNA sequencing capacity, their systems will be rendered obsolete within two years, when a single machine will have more capacity than their combined ~$90M investment has given them in 2010. However, if this existing capacity, purchased at a premium cost, gives them an opportunity to more quickly make valuable correlations between genotype and phenotype, or if this capacity allows them to become the established market leader in a fee-for-service operation, or if this data allows them to make significant strides in pharmaceutical development, then the investment may have been most worthwhile.

4 THE ERA OF PERSONAL GENOMICS HAS ARRIVED

Given the initial high cost, among the first individuals to have their genomes sequenced have been heads of human genome sequencing teams, or researchers who developed their own sequencing systems. Given the striking pace of DNA sequencing methods development, well-funded genome center laboratories are currently acting on plans to sequence thousands of humans and other organisms whose genomes are equally complex. Even at this time, the cost of collecting the genomic data for an individual is within reach to the curious or to patients whose physician believes that a genome sequence may help with disease diagnosis or treatment. Illumina is using their HiSeq 2000 machines to sequence human genomes for a fee, but has provided sequence data for free in special medical cases.

Such requests for complete genome sequence data collection would remain in the realm of the well-funded laboratory or of the wealthy without the development of third-generation DNA sequencing systems. However, as described in Section 3 above, the cost, speed and accuracy of DNA sequence gathering is certain to improve rapidly. At a retail cost of under $500, collection of an individual's genome sequence becomes similar in cost to a typical medical diagnostic ordered by a physician. When this cost threshold is reached, millions of patient's genomes could be sequenced if associated criteria for large-scale human genome sequencing are also met. These associated issues are briefly noted below.

DNA sequence data versus genome information. It is important to note that DNA sequencing systems only generate sequence data. Interpreting the meaning of this data to make appropriate medical decisions involves non-trivial data processing steps including accurate genome reassembly and predictions of phenotypes based on this sequence data. For example, reassembling raw sequence reads to form complete chromosome-size reconstructions is complicated by the presence of repetitive regions of the human genome. Also, any errors in sequence reads that are not eliminated by employing multiple read coverage of each nucleotide may be interpreted as single-nucleotide polymorphisms (SNPs), which may lead to a false prediction of phenotype. Of greater concern is the dearth of knowledge on the impact of genetic differences on many disease or normal phenotypes. These topics will be addressed in greater detail in later sections.

DNA arrays versus DNA sequencing. DNA arrays use immobilized DNAs as hybridization probes to selectively bind to DNA or RNA sequences in a biological sample. The arrays are limited in that they only can detect the sequences they are designed to target. Thus, DNA arrays

are usually designed to selectively report the presence of specific sequences that have proven biological meaning, such as SNPs that are indicative of disease phenotypes. In contrast, complete genome sequencing will report the presence of any sequence signatures, including those whose biological implications are as yet unknown. The collection of complete personal genomes on a large scale could be limited if DNA arrays can be used to gather the most pertinent information more cheaply.

In the past, DNA arrays have been a far more economical platform to gather large SNP and other genetic diversity data sets. However, they are less well suited to identify novel DNA variation that would be useful to diagnose rare genetic diseases or to carry out exploratory studies to identify new SNP-phenotype correlations. Regardless, the rapid reductions in genome sequencing costs will likely permit users of genetic variation information to collect all sequence data for the same or less cost than collecting a partial set with DNA chip systems. Therefore, complete genome sequencing will likely soon replace DNA chips for the collection of DNA variation data in large-scale exploratory studies. DNA chip systems are likely to be used only when precise and well-validated genetic variation is sought and if the DNA chip platform is lower in cost compared to complete genome sequencing.

Correlating human genetic variation with phenotypes. Genetic variation such as SNPs can alter the sequence of proteins by changing the codons of a gene. These changes can be highly predictable based on our understanding of the genetic code. However, mutations can also occur in a promoter region of a gene or in key processing signals of the corresponding messenger RNA that can change the amount of protein production. Given the incomplete knowledge of these genetic control elements, the effects of sequence variation in these regions are less predictable. Moreover, not all genes code for proteins, and there is an increasing recognition that large numbers of noncoding RNA transcripts play important roles in human cells (Wilusz et al., 2009). There is no predictive paradigm equivalent to the genetic code for interpreting mutations in the vast regions of the human genome that do not code for proteins.

Even more confounding is that validated effects of genetic variation on protein expression or function do not always lead to a specific phenotype with 100% certainty. Proteins operate in a complex milieu of other proteins, biopolymers and metabolites and so the partial or even complete disruption of an individual protein's function could be offset by the action of a redundant protein or otherwise have imperfect effects that are mitigated by complex cellular adaptations. Furthermore, epigenetic effects wherein phenotypes are derived from factors other than nucleotide sequence appear to be a major determinant of human phenotypes.

It may be more eommon to see SNPs eorrelate imperfeetly with disease or other phenotypes, such that the deteetion of a SNP in an individual ean at best bc used to provide a probability that the person will develop a partieular disease sometime in the future. One of the greatest challenges in developing personal genomies applications will be the identifieation of validated eorrelations between genetic signatures and important phenotypes. Limitations to personal genome sequeneing may be eneountered if the medical eommunity determines that the utility of eomplete genome sequenee data is exeessively hindered by the lack of validated correlations between genetie variation and phenotypes. Aspects of this important area of researeh that bear on phenotype predietions will be discussed in detail in a later seetion.

Infrastructure for large-scale personal genome sequencing. Within the next several years, third-generation DNA sequencing systems likely will be available that allow the widespread sequeneing of personal human genomes. Along with the requirements for numerous sequencing maehines, the spaee to house them, and the reagents and individuals needed to operate them, there are other infrastrueture needs that aeeompany any high-throughput DNA sequeneing operation. Initially, DNA samples must bc collected and stored, at least until the sample has been sequeneed. It may also be advantageous to arehive these DNA samples indefinitely, particularly if long term epigenetics studies are planned.

The demand for eomputing and data storage capacity also will likely bc large. Raw sequencer data (c.g. fluoreseent image data, eonduetanee data, ete...) must be stored until the DNA sequenee reads can be made. DNA fragment scquenee data must be stored until the genome sequenee can be properly assembled. Finally, thc eomplete genome sequenee must be stored at least until the dcsired genctic information has been obtained. Again, it may be advantageous to store genome sequence data for the life of the individual and beyond, and use medieal and performance records to link genetie data to disease and other phenotype data.

Although there are many faects to establishing thc infrastructure neecssary to eonduct genomie data assessments on a scale required for personalized genomics, several companies are beginning to address these needs and are likely to ercate viable data collection and analysis pipelines beginning with sample eollection and ending with a genotypic and phenotypic report that is delivered to customers. The devclopments made by thc commercial efforts in personal genomics eould be leveraged by the DoD to earry out personal genomics analyses on military personnel.

5 CORRELATING GENOTYPE WITH PHENOTYPE

In some instanees, genetie diseases or other human phenotypes will be readily dedueible from personal genome sequenee data. Highly predietable traits typieally follow simple Mendelian inheritanee wherein a single gene is the eausative genetic faetor. For example, siekle eell anemia typieally is eaused by a single readily-deteeted mutation in the β-globin gene. Individuals who carry one mutant eopy and one normal eopy of this gene are resistant to malaria infeetion, whereas two mutant copies of the gene results in siekle eell anemia. A similar trait with simple Mendelian characteristics is susceptibility to ehronic beryllium disease or berylliosis. Certain variants of human major histoeompatibility eomplex (MHC) elass II proteins allow presentation of beryllium to T eells, thus eausing inflammation (Dai et al. 2010).

Unfortunately, most eommon diseases and other phenotypes of interest are not monoallelic, but rather have eomplex multi-faetor origins that are derived from genetie and/or epigenetie faetors or influenees. Thus, predietions of phenotypes based on DNA sequenee and/or epigenetic data will be imperfeet, and provide only a probability estimate that a particular phenotype will manifest. Given these ehallenges, there will be inaccurate assessments of risks based on genotypie markers, and it will take decades of careful researeh to produee highly aeeurate information for most phenotypes of interest.

5.1 Ongoing Efforts to Correlate Genotype with Phenotype

The explosion of available human genome sequenee data will provide researehers from aeademia and industry with the genetie information neeessary to eonduct large-scale efforts to link genetie markers with human traits. Furthermore, the genetie information must be gathered from individuals who have doeumented diseases or other traits. These analyses, commonly ealled "genome-wide assoeiation studies" or "GWAS" do not require any hypothesis regarding the origin or meehanism of the disease or trait, but rather rely on the identifieation of eorrelations between the phenotype and a genetie marker or sets of markers.

Whereas previous efforts to link genes with phenotypes largely have been foeused on addressing a single mutation or a single trait, GWAS studies have begun to foeus on entire genomes and multiple traits. These studies can be eonducted on groups of unrelated individuals,

but there are particular advantages if the individuals are genetically related (Roach et al. 2010). As the number of validated correlations between genetic markers and traits grows, assessment of an individual's genome sequence and its impact on traits will become more informative (Ashley et al. 2010).

Pharmaceutical and academic laboratories, as well as private personal genome companies, have the technical and financial resources necessary to make considerable advances in this area, and it is evident that they are accelerating genomic data collection and related discovery work at an unprecedented pace. Potential major contributors to this effort are personal genome companies. These companies collect biological samples from customers who are curious about their genetic make-up and wish to receive information on their heritage and various traits. These companies conduct genotype analyses (for example by using gene chips that report the presence of single-nucleotide polymorphisms), and then report on the known correlations between these SNPs and various diseases or other traits. Interestingly, private citizens appear to be willing to fund this data collection, as well as to provide personal health and trait information that may be difficult for studies conducted by pharmaceutical companies or academic scientists. This willingness of the general public, aided by internet-based organizational systems, has yielded novel genetic correlations with human traits in a surprisingly efficient process (Eriksson et al. 2010). Despite these impending advances, additional factors noted in the next section will impede the full assessment of human traits.

Given these advances and limitations, several major outcomes are likely to occur:

- New genetic links to many phenotypes in addition to disease phenotypes will be revealed at a rapid pace by industry and by academia, perhaps using entirely new models for conducting GWAS studies.
- Many of the new discoveries will be revealed to individuals over the internet on a fee-for-service basis, with companies and customers for their information spread across the globe.
- The quality of the predictive information will be mixed, and is likely to be particularly poor for those traits that have complex origins and many contributing factors.
- Given the current state of technology (e.g. the lack of inexpensive access to a complete dataset for the human epigenome), there will be many phenotypes that may be difficult to predict based solely on easily obtainable genomic and epigenomic data.

5.2 Issues that Confound Genotype/Phenotype Correlations

Epigenetics, beyond DNA sequence. Although the term "epigenetics" was first used more than 60 years ago before the basic molecular details of inheritance were elucidated, epigenetics today refers to the study of heritable traits that arise from changes *other than* changes in DNA sequence (Jirtle and Skinner, 2007). It is now known that many key biological features inherited from parent cell to daughter cell or from parent organism to progeny, including those relevant to human disease, are mediated by mechanisms that do not involve simple changes in genome sequence. While interpreting the effects of DNA sequence variation on the genetic code is simple and highly predictable, an "epigenetic code" is currently lacking and therefore interpreting the effects of epigenetic changes cannot yet be assisted by a simple set of rules. Given the importance of epigenetic effects on disease and other phenotypes, scientists are working to reveal the molecular basis of epigenetic effects on living systems. Some of the above-described technologies and approaches to analyze the human genome are also applicable to the human "epigenome" (the complete collection of epigenetic changes in an organism). Analogously, the characterization and analysis of all epigenetic changes in an organism represents the relatively new field of "epigenomics".

By definition, the effects of epigenetic changes do not arise from modifications to the DNA sequence of an organism. Accordingly, epigenetic effects largely arise from differences in the way that a given gene is expressed, rather than differences in its sequence. Different cells within the same organism, and even the same cell at different points in time, can express dramatically different sets of genes. This fact helps to explain why liver cells and brain cells behave so differently despite their virtually identical genomes, and why cells in a developing human embryo have very different properties than cells in an adult.

In addition to arising during the natural course of cellular development and differentiation, epigenetic changes are also thought to arise from environmental conditions, treatment with drugs, aging, diet, and disease. As a result, the ability to characterize the complete epigenome of a human has the potential to shed new light on the biological status of an individual as well as suggest ways of improving the individual's health, not unlike the impact of revealing an individual's complete genome sequence.

It should be emphasized, however, that while connections between genotype and phenotype that are essential to reaping the benefits of human genome sequence data have recently begun to be realized, connections between a person's epigenome and phenotype are significantly less well

established, even though such connections are widely anticipated to exist. Moreover, in contrast to a person's genome which, to a first approximation, is identical between nearly all cells in the body, the human epigenome is known to vary widely between different cell types and to change over time. The cell-specific and dynamic nature of the human epigenome greatly complicates efforts to systematically collect and make use of epigenomic information. These challenges underlie the relative infancy of epigenomics compared with genomics.

Nevertheless, epigenomic data currently can provide valuable and occasionally useful insights into human biology and disease. Knowledge of the genome-wide pattern of epigenetic marks is currently used as a measure of which genes and which regions of the genome are being actively expressed or silenced. Certain patterns of epigenetic change have been connected to developmental steps during embryogenesis and can serve as hallmarks of cellular differentiation and dedifferentiation. Some human diseases are thought to have epigenetic origins, as evidenced by dependence of the occurrence of certain rare human diseases on the epigenetic state of maternal or paternal DNA ("genomic imprinting"). In addition, some carcinogens have not been observed to induce DNA mutations and thus may be increasing cancer risk through epigenetic mechanisms. Finally, teratogens (compounds that induce developmental abnormalities) have been observed to induce effects in generations beyond that of the exposed individual, consistent with an epigenomic effect.

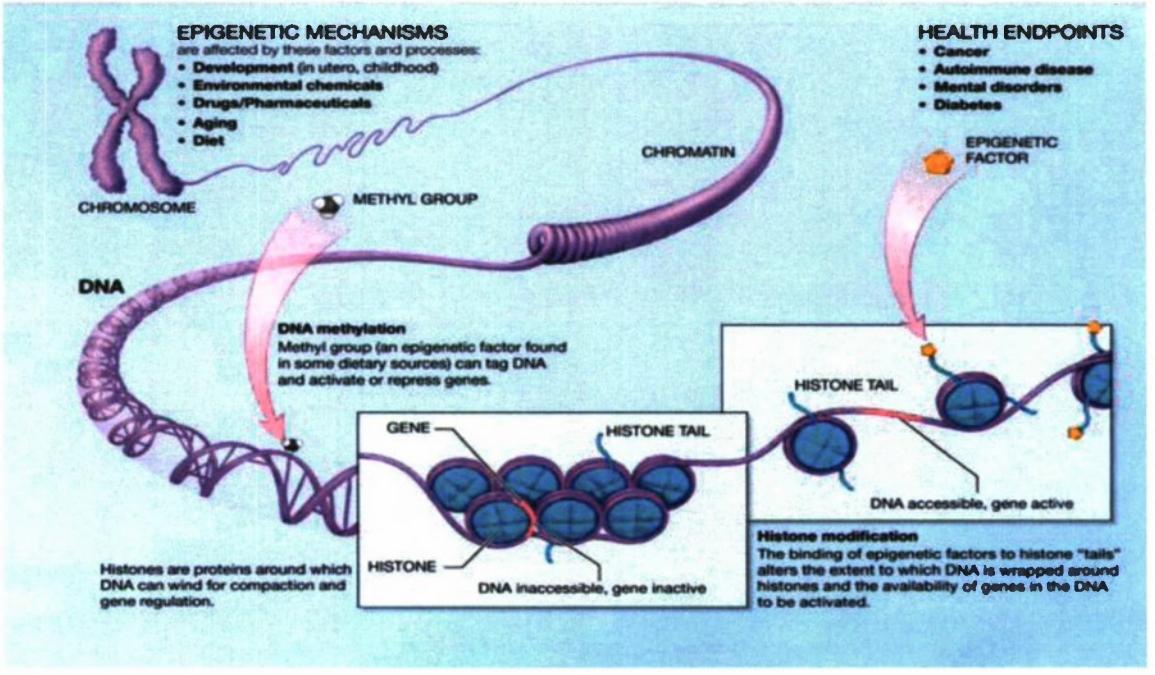
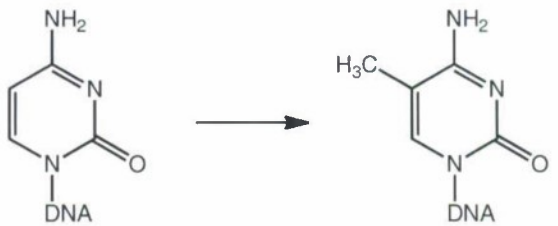


Figure 8. Major epigenetic mechanisms involve DNA methylation, histone positioning, and histone modification.

At the molecular level, most known epigenetic effects arise from modifications in the structure of DNA or the structure of proteins that influence gene expression. However, the known collection of possible epigenetic changes ("marks") in a human is expanding (**Fig. 8**). Currently, these changes can be classified into four major categories: (1) methylation of bases in genomic DNA; (2) post-translational modification of histone proteins and histone positioning; (3) RNA-mediated changes in gene expression; (4) Prion proteins. The first two categories of epigenetic marks are much better understood and have been more widely observed than the last two. The molecular characteristics of these first two classes of epigenetic changes, their known biological roles, and the technologies used to detect these marks are summarized below.

DNA methylation. A common type of epigenetic mark that can influence gene expression is the methylation of cytosine bases in genomic DNA, primarily in cytosines that precede guanine (CG dinucleotides), to produce 5-methylcytosine (**Fig. 9**). Cytosine methylation and CG dinucleotides in general have been observed throughout the human genome, and are underrepresented within the protein-coding regions of genes. It is estimated that 60-90% of all CG dinucleotides are methylated in mammals. DNA methylation is a crucial part of embryonic development, cellular differentiation, and X-chromosome inactivation. The process is also thought to play roles in suppressing the expression of genome parasites such as some viruses and transposons. In some cancers tumor suppressor genes are silenced by DNA methylation, contributing to oncogenesis.



cytosine 5-methylcytosine

Figure 9. Chemical structures of cytosine and 5-methylcytosine.

Newly synthesized DNA generated by DNA replication is unmethylated. The transfer of a methyl group to cytosine is mediated by a class of enzymes called DNA methyltransferases and several of these enzymes have been characterized in humans. The properties of one DNA methyltransferase in particular, DNMT1, demonstrates one way in which epigenetic marks can be inherited. DNMT1 binds especially well to replicating regions of DNA that are hemimethylated (methylated on one strand only) and it is thought that by preferentially binding

to hemimethylated DNA, DNMT1 methylates the newly synthesized, non-methylated strand and thereby copies the methylation marks from a parent cell to a daughter cell. It is also known that the methylation state of DNA in germ line cells can persist in newly fertilized eggs.

Regions of DNA that contain methylated cytosines generally are less transcriptionally active and suppressed in gene expression. The mechanism of this suppression is not entirely understood, but methylcytosine binding proteins that are attracted to 5-methylcytosine and promote condensation of nearby DNA into transcriptionally silent chromatin have been implicated in the process. In addition, some protein transcription factors that are necessary for the expression of certain genes are thought to bind more poorly to methylated DNA.

Several techniques have been developed to detect methylated DNA. For example, certain restriction endonuclease enzymes will only cleave DNA containing methylated, or unmethylated, cytosines. Antibodies also exist that can bind specifically to methylated, but not unmethylated, DNA, enabling fragments of methylated DNA to be isolated and then identified by binding to specific sequences present in DNA microarrays.

Techniques for methylated DNA that can be applied in a genome-wide manner are of particular interest. Sodium bisulfite reacts with unmethylated cytosines, but not methylated cytosines, to produce uracil. Most DNA sequencing methods including those described earlier in this report read uracil as thymine (T). As a result, comparing the sequence of DNA before and after bisulfite treatment can reveal unmethylated and methylated cytosines on a genome-wide scale with a speed and cost that is near equivalent to that for conventional DNA sequencing. Therefore, it is reasonable to expect that large-scale personal epigenomics data will be collected in the future. However, it is important to note that DNA methylation patterns vary from cell to cell, change over time, and are affected by environmental conditions. Thus it is currently impractical to collect a complete DNA methylation epigenome data set at this time.

Histone protein modification. Covalent modifications to histone proteins represent the second major class of epigenetic changes. Histones are highly positively charged proteins that serve as molecular spools around which highly negatively charged DNA is wound and organized to form nucleosomes (**Fig. 10**). Covalent modifications to histones that have been observed in cells include methylation, acetylation, and phosphorylation, among other changes. These modifications are especially prevalent at the N-termini of histone proteins, known as the "histone tails". The most well-characterized of histone modifications are acetylation and methylation of lysine residues within histone proteins.

Acetylation of histone lysines is thought to result in the activation of gene expression. Simplistically, lysine acetylation transforms a positively charged lysine residue into a neutral acetyllysine residue. As a result, histone proteins that are highly positively charged before acetylation become less positively charged and are less strongly attracted to DNA. This weakening of the histone protein-DNA attraction liberates DNA from its previously bound state and facilitates binding of transcription factors to DNA, stimulates transcription, and increases gene expression. Conversely, histone deacetylation can repress gene expression. Methylation of histone lysine residues, in contrast, has been observed to either activate or suppress gene expression, depending on which histone lysine residue is being methylated and even depending on the location of the gene relative to the histone protein. Some of the enzymes that catalyze the acetylation, methylation, deacetylation, and demethylation of histone proteins have been identified and characterized.

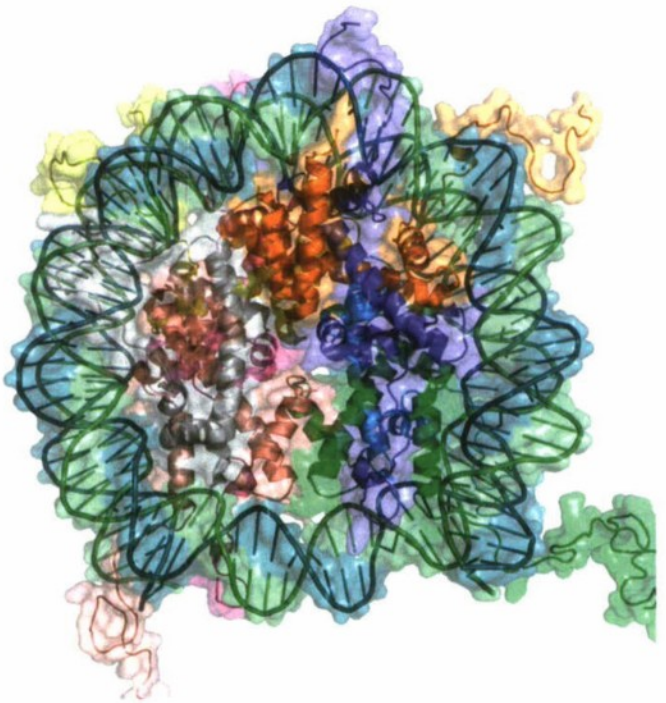


Figure 10. A nucleosome structure formed by DNA wrapped around a core of histone proteins. The positions of nucleosomes in genomic DNA and the modification state of the associated histones influence gene expression.

Since DNA can retain its associated histone proteins after DNA replication takes place, the modification state of DNA-associated histones can be inherited from parent cell to daughter cell after mitosis. In a manner analogous to the action of DNMT1, it has been suggested that the presence of modified histones on a newly replicated genome can induce the modification of other

nearby histones, providing an additional mechanism by which the modification state of histone proteins can be passed on during cell division.

The most common methods to detect histone modifications rely on antibodies raised to bind specifically to methylated (or acetylated, non-methylated, or non-acetylated) histone proteins. In one popular suite of techniques known as chromatin immunoprecipitation (ChIP), genomic DNA with bound histones is fragmented, and fragments containing specific histone modifications are isolated by precipitation with the abovementioned antibodies. The precipitated DNA fragments are inferred to be bound to histone proteins with the specific modification of interest, and are identified in a genome-wide manner by exposure to a DNA microarray or by high-throughput DNA sequencing, or in a gene-specific targeted manner by PCR. As with all antibody-mediated techniques for identifying epigenetic changes, the quality of the resulting data is strongly dependent on the binding specificity and affinity of the antibodies in use, many of which are of questionable quality.

Because the modification state of histone proteins determines their ability to repress or activate gene expression, a wide variety of biological processes including those central to human development, responses of human cells to environmental cues, and changes in gene expression associated with many human diseases are thought to be impacted by histone modification. However, the diversity of modifications, the variable sites and extents of these modifications, and the different effects these modifications can have on chromatin structure and gene expression are confounding aspects of epigenetics that preclude the rapid collection of a complete set of these epigenetic marks.

6 MICROBIOMICS – THE IMPORTANCE OF MICROBES ON HUMAN PHENOTYPES

Another influence on human phenotypes that is not accessible by evaluating personal genomic data is the human microbiome. The diversity of microbial species associated with each person is vast, distinctive and highly variable. However, until recently, researchers have mostly ignored these organisms unless they are the direct causative agents of bacterial or viral infections. NIH has established the Human Microbiome Project to "generate resources enabling the comprehensive characterization of human microbiota and analysis of its role in human health and disease". Such research efforts should provide a clearer understanding of the types of organisms that are present under normal circumstances, and provide a sense of the effects of these organisms on human phenotypes, including disease.

6.1 The Diversity and Scale of a Personal Human Microbiome

The microbial flora in a human gut (~10^{14} cells) and on human skin outnumbers the human cells in your body (~10^{13}) by over 10 to 1. As many as 10,000 distinct species of microorganisms and viruses can be associated with each person. Although bacterial genomes have about an order of magnitude fewer genes per genome, the great diversity of microbial species associated with humans means that the diversity of microbial gene functions may exceed that of human DNA by 100 fold.

Traditionally, the medical community has been exclusively focused on viral or bacterial species that are clearly linked with infections. Bacterial pathogens or their disease phenotypes typically are diagnosed and an appropriate course of antibacterial therapeutics is delivered to kill the infectious organism. However, researchers are now beginning gather data suggesting that some common human phenotypes, such as obesity or drug tolerance/efficacy, are influenced by the microbes associated with each person.

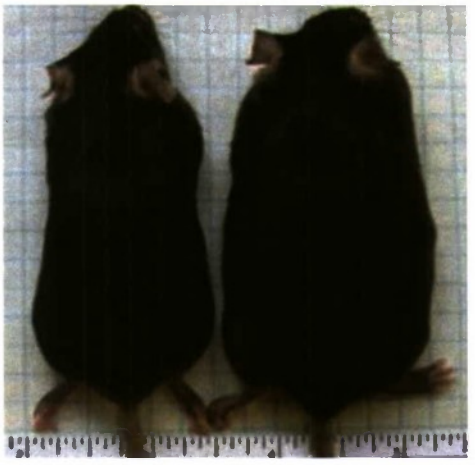


Figure 11. Images of mice with gut microbes for thin (left) versus obese (right) delivered by fecal transplant. Image from Vijay-Kumar et al. 2010.

6.2 Phenotypes Influenced by the Microbiome

In a recent study (Vijay-Kumar et al. 2010), gut microbes from obese mice were transferred to thin mice whose gut microbes were eliminated. This "fecal transfusion" resulted in the thin mice becoming obese (Fig. 11), demonstrating that the microbial composition of the gut can dramatically affect calorie uptake, most likely by affecting appetite. A similar "fecal therapy" approach is used in some instances of *Clostridium difficile* infection in humans (Schwan et al. 1983). *C. difficile* may normally reside in an individual's gut, but rarely causes disease when a typical distribution of microbes is also present. However, in cases where the normal bacterial species have been reduced (*e.g.* during treatments with antibiotics), *C. difficile* becomes the dominant species, produces toxin that damages intestinal cells, and can eventually cause death. The infection is usually treated with antibiotics, but drug resistant strains can emerge. However, the transfer of a fecal sample from a healthy individual to a patient infected with *C. difficile* can repopulate the gut with a normal spectrum of microbes and cure the symptoms of *C. difficile* infection.

An area of great potential involves the assessment of the effects of microbes on the efficacy or toxicity of therapeutic drugs. Given the extraordinary genetic diversity of the gut microbiome, it is expected that these organisms (which usually have the first chance to interact with orally bioavailable drugs) may express protein enzymes that catalyze metabolic changes to compounds that render them inactive or that convert them into toxic derivatives. If true, then metagenomic analysis of gut microbes could be used to identify organisms that could cause drugs to be inactive or toxic, perhaps allowing their selective elimination via targeted antibiotics treatment, followed by delivery of the drug to the patient.

6.3 Potential Uses of Microbiome Data

Individual actions (diet, therapeutic drug use, recent travels) can dramatically influence the distribution of microbial species that are associated with each person. This great and ever-changing diversity of commensal microbial species is not evaluated by genomics efforts that focus on human genome sequence alone. Therefore, assessment of the microbiome of an individual, for example by isolating microbial DNA and conducting metagenomic sequence analysis, could be useful for a variety of purposes.

The distribution of microbial species changes dramatically when individuals are treated with broad-spectrum antibiotics to favor bacteria that resist the antimicrobial action of the drug. Also, the species and their abundances will change in response to the types of food consumed, as certain bacteria may have metabolic pathways allowing them to better exploit the types of sugars or other nutrients in the diet of their host. This remodeling of the microbial species of an individual could be exploited as a new form of tracking, tagging and locating technology that exploits the unique microbiome of each individual. A recent history of drug use, foods consumed, or locations visited could be predicted based on the distribution of bacterial species on skin or in feces. Analysis of microbiome data could also be used diagnostically, by searching for genetic signatures of pathogens. Moreover, traces of a person's microbiome signature will be left on objects they have come in contact with, such as computer keyboards and door handles, just as they may leave traces of their own DNA on these objects.

Microbiome signatures of entire populations could also provide useful information. For example, the identities or unusual microbial genotypes of small populations could be assessed by sampling in vehicles, buildings, or the effluent of entire villages. The latter example could be scaled up to assess effluent at specific nodes of a sewer system of a major city potentially to track individuals to specific locations in the city via their unique genomic or metagenomic signatures deposited in sewer systems. Perhaps a more useful version of this approach would be to create a "disease weather map" that correlates temporal changes in pathogen metagenomics with geography.

7 GENOME DATA STORAGE, ANALYSIS AND SECURITY

As the plans for gathering large numbers of personal human genomes expand, there is a need to consider the needs for storage, analysis and security of the resulting data and information. Very quickly, it becomes apparent that the costs of these necessary nodes in the pipeline of personal genome information become limiting relative to the costs of DNA sequencing. Any future large-scale personal genomics effort will need to address a variety of factors related to the storage, analysis and security of genomics information.

7.1 Computational Costs for Genome Data Storage and Analysis

As the costs of DNA sequence data collection falls, the costs of other needs associated with personalized genomics become a greater portion of the total expense. Of particular concern is the management and analysis of human genomic data (Richter and Sexton, 2009). The computational costs of a single human genome (**Table 1**) can be measured in dollars, cpu-hours, storage, bandwidth or human analyst time. The example below of sequencing of one human genome using the Illumina HiSeq 2000 system provides a clear understanding of why this aspect of personalized genomics will exceed that of DNA sequence data collection. This example considers just the results and cost of initial data processing. Longer term analysis and storage of the data will of course increase the total cost.

Table 1. Costs related to the storage and processing of DNA sequence data from a single human genome. Question marks designate unknown parameters.

Data Type	Format	Compute Time	Compute Cost	Data Size	Storage Cost/Year	Transfer Time	Transfer Cost
Raw Image Files	.tiff	n.a.	n.a.	30 TB	$36,000	18 days	$3,000
Unassembled Reads	.bcl	n.a.	n.a.	100 GB	$120	1.4 hr	$10
Mapped Reads	.bam	500 cpu-hr	$50	100 GB	$120	1.4 hr	$10
Assembled Genome	.fasta	1500 cpu-hr	$150	6 GB	$7	5 min	$0.60
Differences from Reference	?	?	?	4 MB	$0.005	0.2 sec	$0.0004

Raw data generation. An eight-day run of one Illumina Hi-Seq 2000 machine generates about 200 gigabases of DNA sequence consisting of 2 billion reads, each about 100 nucleotides long. As raw data, the run produces a total of about 60 terabytes (TB) of TIFF image files. This large amount of data cannot be efficiently stored, and therefore the system processes the image

data to generate DNA sequence data ("base calls") plus a record of the data quality (a code indicating the software's estimated error probability for each base call). Illumina's base-calling software produces binary files (called .bcl files) containing this information, requiring 1 byte/base and a total of 200 gigabits (GB) of base-called sequence reads.

Base-calling is done in near-real-time using built-in software and hardware that is included with the machine, so no extra computational costs are assigned to this step. 200 gigabases is sufficient to cover two human genomes at 30-fold coverage. 30-fold coverage is approximately the minimum required to reasonably reconstruct whole-genome sequence data via shotgun sequencing. Therefore it is estimated that, for one human genome, 30 TB of raw image data and 100 GB are required for one billion unassembled sequence reads including quality values.

Read mapping. Many analysis tasks involve detecting small differences (substitutions and small insertions/deletions) relative to a reference human genome. For this, unassembled reads from a new genome sequence data set only need to be mapped (aligned) to the reference. The program BWA, an example of the current state of the art in read mapping (using an algorithm called the Burrows/Wheeler transform), can map one billion reads in about 500 cpu-hours (Li and Durbin, 2009). Mapped reads are stored in a binary format called BAM (Li et al., 2009), requiring 1 byte/base and a total of 100 GB for 1 billion reads.

Whole-genome shotgun assembly. Tasks such as studying large-scale rearrangements and structural differences may require de novo shotgun assembly to create a contiguous whole genome, which is a harder problem than just mapping reads to a reference genome. Currently, assembly algorithms for new sequencing platforms are immature (Miller et al. 2010). There are as yet no examples of successful de novo assembly of highly accurate, highly contiguous mammalian genomes from Illumina sequencing data. An example of the state of the art is the program SOAPdenovo, which has been reported to assemble a human genome from 52-fold coverage (Illumina) in 1500 cpu-hr on a 32-core, 512GB RAM system, resulting in 80% genome coverage in contigs (continuous overlapping assembly of DNA) of average size 1 kb (Li et al., 2010). An unannotated diploid human genome assembly is typically stored in FASTA (text) format, requiring about 1 byte/base and a total of about 6 GB for one diploid genome.

Compression. If a large number of assembled genomes need to be stored, this can be done efficiently by only storing differences relative to a reference genome. A recent proof-of-concept paper showed that a single human genome can be stored in 4 MB, using a 3 GB reference genome and a 1.2 GB reference table of the most common single nucleotide polymorphisms (Christley et al., 2009). Since the compute costs were not provided, this is shown as unknown

(**Table 1**), but are likely to be less than read mapping or shotgun assembly. However, compression of this sort is not yet in widespread use. Current sequence analysis software still generally assumes sequence files in FASTA format. The 1000 Genomes Project is currently working to standardize a system for compact representation and storage of human genome variation. This representation (or something like it) can be expeeted to percolate into downstream analysis software, as pressure from human genome data volume increases.

Compute cost. Computing cost is estimated at about $0.10/epu-hr, based on two data points. The Amazon.com EC2 compute cloud charges between $0.085 and $2.88 per cpu-hr. The fully loaded cost of the Howard Hughes Medical Institute Janelia Farm computing cluster, an example of a midsized (4096-core) scientific computing resource for biological research (including power, infrastructure, amortized capital expenses, and staffing) is $0.11/cpu-hr .

Storage cost. Storage cost is about $0.10/GB/month, based on two data points. Cloud storage (for example Amazon.eom S3) is currently charged at $0.055-$0.150/GB/month. The fully-loaded cost of primary high-availability storage on the Janelia Farm computing resource is $0.18/GB/month and long-term archival storage is $0.06/GB/month, both ineluding onsite backup and offsite disaster recovery.

Transfer cost (Internet). Network data transfer cost is estimated at about $0.10/GB, based on two data points. A 155 megabit OC3 connection (maximum bandwidth 50 TB/month) costs about $7500-$40,000 per month, depending on the Internet service provider, which is approximately $0.40 - $2/GB. Data transfer out of the Amazon.com S3 cloud is priced at $0.08 - $0.15/GB.

Transfer time. Transfer time is calculated based on a dedicated 155 megabit/sec OC3 connection, to give a reasonably "typical" example for a research center. In practice, available bandwidth varies quite widely. Many current Internet nodes would have less than OC3 connectivity. For example, a typical home connection might have 1-10 megabit/sec download. Some sites could have much more than OC3 connectivity. One example of a very high end exeeption is the National Scienee Foundation Teragrid, with participating national Teragrid sites interconneeted at 10 GB/see (about 70 fold faster than our chosen example). The cost and time required for networked data transfer for large data sets ean be prohibitive. Large data sets are typieally transferred by shipping disk drives. The Amazon eloud, for example, has an efficient data ingest system based on overnight FedEx shipping of disks.

7.2 Combining Personal Genome Data with Existing Health Records System

Electronic medical record keeping in the military. The Defense Health Information Management Systems (DHIMS) oversees the military's electronic health record (EHR) system, which contains the medical records of nearly 10 million military service personnel and their families. The first version of this EHR system, named Composite Health Care System I (CHCS I), was developed by SAIC and implemented in the late 1980s. It was based on the existing Veterans Health Information System and Technology Architecture (VistA). CHCS I relied on terminals with a command line interface, connected to a host computer at individual hospitals. The second phase of this system, CHCS II, was implemented in the late 1990s and enabled data sharing among military health care facilities. In 2004, CHCS II became the Armed Forces Health Longitudinal Technology Application (AHLTA), which has a graphical user interface and provides more comprehensive access to medical information. AHLTA contains dozens of individual modules, covering topics such as medical history, patient visits, medical orders, prescriptions, and laboratory test results. It captures approximately 100,000 patient encounters per day, both in CONUS and for those deployed overseas.

AHLTA, currently in version 3.3.3, has become somewhat unwieldy, often requiring painstaking data entry procedures. The current version of the manual contains 631 pages, and proper use of the system necessitates classroom or web-based training. ALTA is used at nearly 500 military treatment facilities worldwide. There also are special versions of the system, such as AHLTA-Theater (for outpatient facilities in theater), AHLTA-Garrison (for theater hospitals, forward resuscitative sites, and U.S. Navy ships), and AHLTA-Mobile (for battlefield use). Even the White House Medical Unit uses AHLTA.

The Veterans Health Administration (VHA) continues to rely on VistA as its EHR system. The VHA is the largest single medical system in the U.S., providing care to over 4 million veterans, and operating 163 hospitals and nearly 1,000 clinics and nursing homes. VistA is organized based on patient encounters at individual treatment centers, rather than aggregated patient-specific information, but is currently undergoing modernization to become more centralized. VistA consists of a suite of individual applications, rather than modules within a common application.

Although AHLTA and VistA are related by descent, the two systems have diverged to the extent that they are no longer compatible. AHLTA must operate under a broad range of environments reflecting a command-and-control organization, while VistA serves a single, very

large civilian health care provider. Some have advocated merging the two systems, based on either of the two architectures, but this is appears unlikely in the near future and will become increasingly difficult as the two systems continue to diverge.

Inclusion of genotype and phenotype data. It would be possible to include genomic and epigenomic data within the existing DoD and VHA EHR systems. This would involve a separate module within AHLTA or a separate application within VistA. Site-to-site variation in the current VistA file structure is a challenge, but presumably will no longer an obstacle once the more centralized version of VistA becomes available. The size of genomic/epigenomic datasets exceeds the largest datasets currently in these EHR systems. The AHLTA Radiology module and VistA Imaging application, for example, include megabyte-sized radiological images, but most of the >50 terabytes of data in these systems are text information. Genomic/epigenomic data might be stored as differences relative to a reference dataset rather than complete datasets for each individual, thereby reducing data storage requirements to the level already implemented in the EHR systems.

Both AHLTA and VistA currently provide a wealth of phenotypic data. This information is organized in relation to the flow of patient encounters and treatments, rather than biological phenotypes. Nonetheless, it should be possible to mine the stored information to conduct hypothesis testing in relation to collected genomic/epigenomic data. Such data mining tools already are being applied to the analysis of clinical outcomes (Moody, 2007).

There are substantial efforts within the research community and various health care delivery organizations to correlate genotypic and phenotypic data for the potential improvement of human health. The DoD should not duplicate these efforts, but rather seek to leverage them in addressing issues that are of special concern to the military. The first step, therefore, is for the DoD to determine which phenotypes that might reasonably be expected to have a genetic component have special relevance to military performance and medical cost containment. These phenotypes might pertain to short- and long-term medical readiness, physical and mental performance, and response to drugs, vaccines, and various environmental exposures, all of which will have different features in a military context. More specifically, one might wish to know about phenotypic responses to battlefield stress, including post-traumatic stress disorder, the ability to tolerate conditions of sleep deprivation, dehydration, or prolonged exposure to heat, cold, or high altitude, or the susceptibility to traumatic bone fracture, prolonged bleeding, or slow wound healing.

7.3 Implications of Computational Costs

The computational costs of even the initial, highly automated data processing pipeline needed to produce sequence data suitable to extract biological information is already nearly equivalent to the costs of DNA sequencing. Because DNA sequencing costs are dropping much faster than computational costs (in cpu-hrs, storage, or bandwidth), computational cost will soon dominate the costs of personal genomics data collection and analysis. Storing the actual raw data from second-generation sequencing platforms is essentially already out of the question. The "raw data" that is more feasible to store are the unassembled reads with quality values, after a computational base-calling step has already been done by the sequencing machine. In addition, any costs will be increased by the need to store the data securely, derive useful information from this data, and eventually couple this information with other data sets such as medical records.

Even greater challenges lie in downstream computational analyses of the data. The National Human Genome Research Institute (NHGRI) of the National Institutes of Health (NIH) is currently developing its latest strategic plan for the next five years. A large component of this plan will be devoted to bioinformatics. Because of the rapid advent of high-throughput data technologies including DNA sequencing, biology is rapidly being transformed into a data-intensive computational science. This raises a number of challenges throughout the biomedical research enterprise, including: training biologists in computational analysis; supporting research teams of biologists and computational scientists; establishing computational hardware infrastructure in biology departments; developing robust analysis software in largely open-source, academic environments; designing and managing strong experimental plans for "big science" teams in biology that better integrate data generation with data analysis; sharing and distributing electronically-readable datasets. Any large research effort in genome sequence analysis will face these same issues, and would probably benefit by cooperating or allying with the NIH NHGRI, which is probably the leading federal agency in large scale data analysis in genomics.

8 PERSONAL GENOMICS DATA AND INFORMATION USEFUL FOR MEDICINE AND DEFENSE

8.1 What Personal Genomics Data should be Collected

Collection of a diversity of personal genetics data would be useful for diagnostic and predictive applications. A summary of the key data sets that would be useful are given below.

Complete Genome Sequences. The complete diploid genome sequences for all military personnel should be collected. This can be used to establish the following genetic signatures:

- Define all the SNPs, InDels (nucleotide insertions and deletions), and copy number variances. This information permits correlations to be made between genotypes and phenotypes of interest to the DoD.

- Establish MHC allelic diversity. This information permits the prediction of donor-recipient pairs and should be predictive of a person's ability to respond robustly to vaccines and new infectious disease challenges.

Unaligned DNA Sequences. Some DNA sequence reads cannot be assembled as part of the human genome, and these sequences are usually discarded. Although the source of some of these reads could be DNA contamination, some reads may represent viral or bacterial pathogens that have infected the individual. Knowledge of the spectrum of pathogens may be of considerable importance.

Personal T-Cell Repertoire. Collecting the DNA sequences corresponding to the unique antibody-coding regions of T-cells may be used to establish allergies and to establish previous exposure (and immunity) to various pathogens.

Human Microbiome. In some cases it may be useful to sequence metagenomic samples of the microbiomes that colonize the human gut, oral cavity or other areas of the body. Uses involve identifying viral or bacterial pathogens, identifying organisms that are not pathogens but are diagnostic for disease or other phenotypes, or can be used as microbiological signatures for tracking, tagging and locating (TTL) applications. The timing of data collection may also be important if seeking evidence of pathogen infections or gut microbiome changes (*e.g.* sequencing both pre and post mission).

Epigenetics Data. Some types of personal epigenetic data may be useful for epigenome-phenotype correlations (e.g. DNA methylation patterns, histone modification patterns). However, there are considerable technical roadblocks that will preclude the collection of a complete epigenome, such as variability between cells, tissues, age, and chemical or environmental exposures.

Data of Importance to the VA. The same data categories listed above should be collected by the VA, but correlations made between the DNA sequence or epigenetic data should be made for phenotypes of interest for post-service care.

8.2 Using Personal Genome Information

In addition to the ethical implications of applying personal genome information, its use must be done with careful consideration of the validity of any correlations between genotype and phenotype. Acting on genotype information that is not convincingly linked to specific phenotypes could lead to erroneous and detrimental decision making. Unfortunately, developing a list of validated correlations is made even more difficult by the complex nature of the underlying genetic sources for some phenotypes. Most common diseases and other phenotypes of interest are not monoallelic, but rather have complex multi-factor origins that are derived from genetic and/or epigenetic factors or influences. This poses the following challenges:

- Predictions of phenotypes based on DNA sequence and/or epigenetic data will be imperfect, and provide only a probability estimate that a particular phenotype will manifest.

- There will not be technologies in the foreseeable future that provide an inexpensive and comprehensive dataset for the human epigenome. Therefore, there may be some important phenotypes that may be difficult to predict based solely on easily obtainable genomic and epigenomic data.

- Given these challenges, there will be inaccurate assessments of risks based on genotypic markers alone, and it will take decades of careful research to produce highly accurate information for most phenotypes of interest.

The human microbiome holds great promise for disease diagnosis, phenotype determinations, and for TTL applications. Microbiomes adapt to the types of food and drugs we ingest and are

usually specific for a person's point of recent habitation. This is potentially a major untapped source of information on a person's health, movements, and recent experiences, and could be used to help determine recent contact between objects and individuals. However, given the transient nature of the spectrum of microbes in personal microbiomes, very little information is available at this time to draw validated correlations between microbiome and health.

In summary, there are many possible useful outcomes to the arrival of the personal genome era, and DoD can position itself to be a leader in areas such as genotype/phenotype correlations, or in human microbiome diversity and its influences on health or its utility in TTL. DoD should pursue its own unique interests in these areas, and also have the infrastructure and expertise in place to position themselves to respond rapidly as new technologies and new findings emerge.

9 CONCLUSIONS AND RECOMMENDATIONS

DNA sequencing is already cheap enough to initiate the era of personal genome sequencing and further reductions in cost will make human genome sequencing increase in scope from hundreds of people (current) to millions of people (probably within three years). Although sequence data collection will not be the rate-limiting step in genome information gathering, there will be substantial developments needed in correlating genotype to phenotype, which is limited by quality phenotype data and computational systems required to make correlations.

Despite these challenges, the DoD and the VA may be uniquely positioned to make great advances in this space. DoD has a large population of possible participants that can provide quality information on phenotype and the necessary DNA samples. The VA has enormous reach-back potential, wherein archived medical records and DNA samples could allow immediate longitudinal studies to be conducted.

Primary Conclusions:

1. The $100 genome is nearly upon us, and soon the cost of DNA sequencing will no longer be a limiting factor in genomic analysis.
2. The era of personal genomics has already begun, but the practical application of genomic information has thus far been limited.
3. Broader application of genetic information will require deeper knowledge of genotype-phenotype correlations, a subject of substantial, ongoing research.
4. Many phenotypes of relevance to the DoD are likely to have a strong genetic component, for which better understanding may lead to improved military capabilities.
5. Certain phenotypes will also depend upon epigenomic and microbiomic contributions. However, human epigenomes and microbiomes are diverse and will change with time, and therefore complete datasets for these genetic signatures cannot be collected.
6. The DoD already maintains a comprehensive medical database for its personnel that eventually will also include their complete genome sequences.
7. The DoD will benefit by organizing personnel data into phenotypes of relevance to the military, then correlating those phenotypes with genetic information.

Recommendations:

The DoD can benefit significantly by employing personal genomics technologies when evaluating the health and performance characteristics of their personnel. The DoD could take a leading role in the personal genomics era, and become full partners with industry and academia in creating useful information from genotype and phenotype data. Alternatively, the DoD could choose to play a more limited role in the research necessary to link genotypes with phenotypes, and pursue only those aspects that are of special interest to the military and that would otherwise not be pursued by the civilian sector.

The DoD can harness the advances in personal genomics technology by taking the actions described below.

Major Recommendation

The DoD should establish policies that result in the collection of genotype and phenotype data, the application of bioinformatics tools to support the health and effectiveness of military personnel, and the resolution of ethical and social issues that arise from these activities. The DoD and the VA should affiliate with or stand up a genotype/phenotype analysis program that addresses their respective needs. Waiting even two years to initiate this process may place them unrecoverably behind in the race for personal genomics information and applications.

Specific Recommendations

DoD Military Health System

1. Establish procedures for the collection and archiving from all military personnel DNA samples that are compatible with subsequent genotype determination.
2. Plan for the eventual collection of complete human genome sequence data from all military personnel.
3. Arrange for the secure, long-term storage of DNA sequence data.
4. Prepare for the collection of epigenome and microbiome data when appropriate.

DoD Office of Health Affairs

5. Determine which phenotypes are of greatest relevance to the DoD.
6. Cooperate with health care professionals to collect and store these data.

7. Use bioinformatics tools to correlate genetic information with phenotypes to discover linkages between the two datasets that will ultimately allow genotype information to be used productively.

References

1. Arnaud, C. H. (2009) DNA sequencing forges ahead. *C&E News* **87**:16-19.

2. Ashley, E. A. et al. (2010) Clinical assessment incorporating a personal genome. *Lancet* **375**:1525-1535.

3. Clark, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S. and Bayley, H. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnol.* **4**:265-270

4. Christley, S., Lu, Y., Li, C., and Xie, X. (2009) Human genomes as email attachments. *Bioinformatics* **25**:274–275.

5. Cyranoski, D. (2010) The sequencing factory. *Nature* 464:22-24.

6. Jirtle, R. L. and Skinner, M. K. (2007) Environmental epigenomics and disease susceptibility. **8**:253-262.

7. Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* **25**:1754–1760.

8. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078–2079.

9. Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J., and Wang, J. (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**:265–272.

10. Li, Y. and Wang, J. (2009) Faster human genome sequencing. *Nat. Biotechnol.* **27**: 820-821.

11. Maxam, A. M. and Gilbert, W. (1977) A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA* **74**:560-564.

12. Miller, J. R., Koren, S., and Sutton, G. (2010) Assembly algorithms for next-generation sequencing data. *Genomics* **95**:315–327.

13. Moody, R. (2007) The AMEDD AHLTA guide to improved healthcare outcomes. AMEDD AHLTA Implementation and Clinical Integration Office, Washington, DC.

14. Riehter, B. G. and Sexton, D. P. (2009) Managing and analyzing next-generation sequence data. *PLoS Comput Biol.* **5**:e1000369.

15. Roach, J.C. et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**:636-639.

16. Sanger, F. and Coulson, A. R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94**:441-448.

17. Schadt, E. E., Turner, S. and Kasarskis, A. (2010) A window into third-generation sequencing. *Human Molec. Genet.* **19**:R227-R240.

18. Schwan, A., Sjolin, S., Trottestam, U. and Aronsson, B. (1983) Relapsing *Clostridium difficile* enterocolitis cured by rectal infusion of homologous faeces. *Lancet* **322**:845.

19. Vijay-Kumar, M., Aitken, J. D., Carvalho, F. A., Cullender, T. C. and Gewirtz, A. T. (2010) Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5. *Science* **328**:228-231.

20. Wilusz, J. E., Sunwoo, H., Spector, D. L. (2009) Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev.* **23**:1494-1504.